A UNIFIED FRAMEWORK FOR GRADUATION

by

R J VERRALL

Actuarial Research Paper No. 91

Department of Actuarial Science and Statistics
City University
London

November 1996

ISBN 1 874 770 99 9

"Any opinions expressed in this paper are my/our own and not necessarily those of my/our employer or anyone else I/we have discussed them with. You must not copy this paper or quote it without my/our permission".

A Unified Framework for Graduation

Dr Richard Verrall[1]

Abstract: This paper presents a unified approach to the main methods of graduation. Included within this framework are parametric graduation (linear and non-linear, including splines), Whittaker graduation, moving-weighted-average graduation, and kernel graduation. The approach is based on the framework of Generalised Linear Models, extended to include non-parametric smoothing. The extended framework is called Generalised Additive Models. The advantages of this approach is that it is possible to perform and compare parametric and non-parametric graduations within the same framework, using the same statistical package, S-PLUS. Also, it suggests possible extensions to the standard theory of graduation theory.

Keywords: Generalised Additive Models, Generalised Linear Models, Graduation.

[1] Department of Actuarial Science and Statistics,
City University,
Northampton Square,
London. EC1V 0HB
email: rjv@city.ac.uk
tel:0171 477 8476

1. Introduction

Classical (non-Bayesian) graduation theory is usually divided into two parts: parametric and non-parametric graduation. Within each of these, there are a variety of seemingly unrelated methods. For example, parametric graduation would include linear models, generalized linear models, non-linear models and splines, and non-parametric graduation would include moving-weighted-averages, Whittaker graduation and kernel smoothing. The purpose of this paper is to bring all of these methods (both parametric and non-parametric graduation) within a unified framework. There are many advantages for this approach. It allows the practitioner to compare and contrast the assumptions made by each method on an equitable basis. It removes the misconception that the many and diverse graduation techniques are completely unrelated. It makes it possible to perform most graduations using a single statistical package, and to compare the results from using different graduation methods, also within a single package. It suggests possible extensions for graduation theory. These include local polynomial smoothing such as the method outlined in section 3.5. A useful reference on local polynomial smoothing is the book by Fan and Gibjels (1996). It is also possible to use non-parametric smoothing methods (in conjunction with parametric graduation) for assessing the goodness-of-fit of a graduation and for preliminary exploration of the data (see section 5).

In this paper, graduation theory will be presented within the general framework of statistical models in S-PLUS. This has the advantages of unifying the theory as mentioned above. Many non-parametric graduation methods are better formulated as Generalized Additive Models (Hastie and Tibshirani, 1990, 1993) than as in existing actuarial theory. Recent papers on parametric graduation have developed unifying theories for graduation, developed modern statistical methods for graduation and shown connections between different methods. For example, Renshaw (1991, 1992, 1995) showed how parametric graduation can be expressed as generalized linear and non-linear models, and how certain of the particular features of mortality data can be modelled. These papers also showed how the package Glim (Francis *et al*, 1993) provides a convenient vehicle for performing parametric graduation. The importance of these papers for a unified framework for graduation is that they expressed mortality models as generalized linear (and non-linear) models. Generalized Additive Models use the same distributional structure, but apply different smoothing techniques (a parametric model can also be viewed as a method of smoothing the data).

Verrall (1993, 1994, 1995) showed how Whittaker graduation can be expressed as a dynamic generalized linear model. Estimation can then be performed recursively, using a method related to the Kalman filter. These papers showed how Whittaker graduation is closely related to parametric methods, and can be regarded as an extension of a particular linear model. They also showed how Whittaker graduation can be placed within the framework of generalized linear models, instead of using a normal approximation to the likelihood which Whittaker (1923) and subsequent authors have used for computational convenience. These papers also showed how many different modelling assumptions can be incorporated, using the same basic philosophy as Whittaker graduation. A different approach to Whittaker graduation is advocated in this paper, in order that the estimation may be performed in the same package as other graduations. This approach utilises natural cubic smoothing splines, and the equivalence between this method and Whittaker graduation is covered in some detail in Green and Silverman (1994).

Gavin *et al* (1993, 1994, 1995) studied the relationship between moving-weighted-average graduation and kernel smoothing. It was shown that kernel methods provide a more satisfactory

approach, and that many recent developments in the statistical literature can be usefully incorporated into the practice of graduation. However, it was also found that other smoothing methods could provide more satisfactory graduations.

Hence this paper will describe how the present theory of graduation can be incorporated and extended within the framework of Generalized Additive Models. The paper is set out as follows. Section 2 contains an outline of the approach to graduation using Generalized Linear Models. Section 3 briefly describes Generalized Additive Models, and shows how the well-known non-parameteric graduation methods can be incorporated.

2. Generalized Parametric Models for Graduation

2.1 Introduction

This section contains an outline of the theory of generalized linear models in the context of graduation. For a fuller exposition, see Renshaw (1991).

In the traditional approach to graduation (not using generalized linear models) weighted least squares estimation, or maximum likelihood estimation with a normal approximation to the likelihood function is sometimes performed using the crude mortality rates. Transformations are also sometimes applied to the crude rates before they are graduated to allow for the fact that it is more likely that the transformed values follow a smooth progression. These considerations are paramount for modelling life tables and are the subject of the seminal paper by Cox (1972). The structure most appropriate for graduation is that of a generalized linear model (Nelder and Wedderburn, 1972). When a generalized linear model is applied there is no need to apply any preliminary transformations to the data. The standard text on generalized linear models is McCullagh and Nelder (1989). Extensive experience in graduation using generalized linear models has been built up in the actuarial literature: see, for example Renshaw (1991), Renshaw and Haberman (1995), Renshaw and Haberman (1996), Renshaw and Hatzopoulos (1996), Renshaw *et al* (1996), Haberman and Renshaw (1996).

In this paper, the structure of generalized linear models will be used, and the smoothing will be extended from solely parametric models to include non-parametric smoothing. In order to introduce the necessary framework, the generalized linear models usually used in graduation will be briefly described before non-parametric smoothing is addressed.

2.2 Models for q_x

Consider the graduation of the probability of death at age x, q_x. This is based on the number of deaths at age x, A_x, and the initial exposed to risk, R_x. Using the usual model for the number of deaths,

$$A_x \sim \text{independent Binomial}(R_x, q_x) \qquad (2.1)$$

and the observed death rate, which is the maximum likelihood estimate of q_x (treating the ages as entirely separate) is

$$\overset{o}{q}_x = \frac{A_x}{R_x} . \qquad (2.2)$$

The likelihood is

$$L(\underline{q}) = \prod_{x=1}^{n} \binom{R_x}{A_x} q_x^{A_x} (1-q_x)^{R_x - A_x}$$

$$\propto \prod_{x=1}^{n} \left(\frac{q_x}{1-q_x}\right)^{A_x} (1-q_x)^{R_x}$$

$$= exp\left[\sum_{x=1}^{n} \left(A_x \log\left(\frac{q_x}{1-q_x}\right) + R_x \log(1-q_x)\right)\right]. \qquad (2.3)$$

This is in the form of a likelihood of a member of an exponential family (see McCullagh and Nelder (1989), for more details of the properties of exponential families). $\log\left(\frac{q_x}{1-q_x}\right)$ is the logit transformation of q_x, and it is natural to apply a parametric model to this. One of the reasons for this is that $\log\left(\frac{q_x}{1-q_x}\right)$ can take any value on the real line, while still constraining q_x to lie in the interval (0,1). It is straightforward to apply parametric models of this within the framework of generalized linear models, and the purpose of this paper is to extend the smoothing to include non-parametric smoothing. This will also be applied to $\log\left(\frac{q_x}{1-q_x}\right)$, using the binomial error structure as before. The natural framework for this is generalized additive models.
The transformation of q_x, usually known as the link function, is denoted by $g(q_x)$. In this case, the link function is the log-odds or logit link function,

$$g(q_x) = \log\left(\frac{q_x}{1-q_x}\right). \qquad (2.4)$$

Note that, in the theory of generalized linear models, it is more usual to define the link function as a function of $E(A_x) = R_x q_x$. This is a trivial reparameterisation, but we do not do it here to avoid confusion with the model for μ_x in section 2.3.
A generalized linear model is applied by equating $g(q_x)$ to a linear predictor (see section 2.4), which is then substituted into the likelihood (2.3). The parameters in the linear predictor are then estimated by maximising the likelihood numerically.

2.3 Models for μ_x

Consider next the graduation of μ_x. To graduate the death rate, central exposures are used, so that in this section R_x denotes central exposures, whereas in section 2.2 R_x denoted initial exposures. The usual model for A_x is

A_x ~ independent Poisson with mean $R_x\,\mu_x$. (2.5)

The likelihood is

$$L(\underline{\mu}) = \prod_{x=1}^{n} \frac{(R_x\mu_x)^{A_x} e^{-R_x\mu_x}}{A_x!}$$

$$\propto \exp\left[\sum_{x=1}^{n}\left(A_x \log R_x + A_x \log \mu_x - R_x\mu_x\right)\right]. \quad (2.6)$$

Again, this is in the form of a member of an exponential family and it is natural to model $\log \mu_x$ rather than μ_x.

Note that the crude rate is $\overset{o}{\mu}_x = \frac{A_x}{R_x}$. The link function in this case is the log link,

$$g(\mu_x) = \log(\mu_x). \quad (2.7)$$

In this model, the linear predictor is usually augmented by a fixed term known as the offset, $\log R_x$. This automatically allows for the exposure term $A_x \log R_x$ in the log-likelihood. An offset is a fixed, known term in the linear predictor which does not contain a parameter to be estimated. It simply corrects for the exposure.

2.4 The Linear Predictor.

Parametric models for graduation generally use a linear (or even non-linear) predictor, η_x. Note that this includes graduation by parametric splines. For example, a straight line model (ignoring the offset) takes the form

$$\eta_x = \beta_0 + \beta_1 x. \quad (2.8)$$

[Note that this is in the simplest form: often x is re-expressed in terms of orthogonal polynomials in order to improve the estimation properties of the model (see below).] The linear predictor is equated to $g(\mu_x)$ or $g(q_x)$ and the parameters in η_x are estimated using the likelihood function as the objective function in an optimisation procedure. In this way, maximum likelihood estimates are obtained. The straight line linear predictor (2.8) can be extended to a polynomial in x,

$$\eta_x = \sum_{j=0}^{k} \beta_j x^j. \quad (2.9)$$

A parametric spline can be defined and inserted into the linear predictor, or non-linear models can be applied, as described by Renshaw (1991).

It is, of course, possible to investigate the use of other link functions from those mentioned in sections 2.2 and 2.3 to further extend the range of models which may be applied. Again, this is discussed in detail by Renshaw (1991).

In common with actuarial graduation practice (see, for example, Forfar *et al*, 1988), we will apply orthogonal polynomials together with age scaling to the age. Thus, throughout this paper, the age, x, has been transformed in the models to $\frac{x-u}{v}$ for suitable values of u and v, and then Chebycheff polynomials have been used.

2.5 Non-Linear Models

Forfar *et al* (1988) defined the "Gompertz-Makeham" and "Logit Gompertz-Makeham" formulae for graduation. A Gompertz-Makeham formula of order *(r,s)* is defined as

$$GM(r,s) = \sum_{i=0}^{r-1} \alpha_i x^i + \exp\left(\sum_{i=0}^{s-1} \beta_i x^i\right),$$

interpreting the sum as *0* if $r=0$ or $s=0$.

Thus, for example, a *GM(2,2)* formula can be written as

$$GM(2,2) = \alpha_0 + \alpha_1 x + \exp(\beta_0 + \beta_1 x).$$

A Logit Gompertz-Makeham formula of order *(r,s)* is defined as

$$LGM(r,s) = \frac{GM(r,s)}{1+GM(r,s)}.$$

Note that if $r=0$ or $s=0$, these can be expressed as linear models. It is usual to use GM formulae for graduating μ and LGM formulae for graduating q, and we consider the likelihood for each of these cases. Consider graduating q_x using LGM*(r,s)*. The log-likelihood (up to an additive constant) can be written as

$$\sum_x A_x \log\left(\frac{q_x}{1-q_x}\right) - R_x \log\left(\frac{1}{1-q_x}\right).$$

From the LGM*(r,s)* formula for q_x, $q_x = \frac{GM(r,s)}{1+GM(r,s)}$,

$$\frac{q_x}{1-q_x} = GM(r,s) \text{ and } \frac{1}{1-q_x} = 1 + GM(r,s) .$$

Thus, the parameters may be estimated by minimising

$$\sum_x R_x \log(1 + GM(r,s)) - A_x \log GM(r,s).$$

For μ_x, the log likelihood is (up to an additive constant)

$$\sum_x A_x \log \mu_x - R_x \mu_x .$$

Hence, if μ_x is graduated using a GM*(r,s)* formula,

$$\mu_x = \text{GM}(r,s),$$

the parameters can be estimated by minimising

$$\sum_x \left[R_x GM(r,s) - A_x \log GM(r,s)\right].$$

These may be minimised using a standard routine in S-PLUS, although some care must be taken with the initial values for the parameters in order to ensure that the optimisation converges to a global minimum.

2.6 Degrees of Freedom.

A well-known concept from parametric regression is the number of degrees of freedom associated with a model, and it will be useful to extend this to non-parametric smoothing methods. For parametric regression, the definition of degrees of freedom is clear: if there are n observations and k parameters in the model, then the model degrees of freedom are k and the degrees of freedom for noise are n-k. To extend this to a non-parametric smoother a different, but equivalent, definition is required. To facilitate this, consider a linear regression model for normally distributed data:

$$Y = X\beta + e$$

where Y is the observation vector of length n,
β is the parameter vector of length k,
X is the design matrix,
and e is a vector of independent, zero mean, homoscedastic, normally distributed random variables.

Then the fitted values are given by

$\hat{Y} = AY$ where $A = X(X'X)^{-1}X'$.

The matrix A is known as the "hat matrix" since it maps the observations to their fitted values. The diagonal elements of A are called the "leverage values" (see Cook and Weisberg, 1982).
The model degrees of freedom are equal to the trace of the matrix A, tr(A), and for this model will be k. The degrees of freedom for noise are tr(I–A), where I is the identity matrix, leading to the same value as above, n-k, for the standard linear regression model.
A similar approach can be taken for non-parametric smoothing, leading to the definition of equivalent degrees of freedom. These are useful for comparing different smoothers, and will be used in section 3. If a non-parametric smoothing method can be written as $\hat{Y} = AY$ for some matrix A, then the equivalent degrees of freedom can be defined as above. In other cases, it is possible to define a suitable quantity or an equivalent linear approximation to the degrees of freedom (see Green and Silverman, 1994).

3. Generalized Additive Models.

3.1 Introduction

In generalized linear models, a linear model is postulated which is equated to $g(\mu_x)$ or $g(q_x)$. For example, a straight line fit could take the form

$$g(q_x) = \beta_0 + \beta_1 x \qquad (3.1)$$

or $$g(\mu_x) = \beta_0 + \beta_1 x. \qquad (3.2)$$

This model is then inserted into the likelihood function and the parameters α and β are estimated by iterative maximum likelihood estimation using a procedure known as Fisher scoring. This is conveniently performed within a statistical package such as Glim (Francis *et al*, 1993). A rich variety of models may be tried, and the statistical basis of these techniques allows a rigorous assessment and validation process to be carried out.
However, generalized linear models are restricted to parametric models. To apply non-parametric models, a relaxation of this method is required. To achieve this, we replace equations (3.1) and (3.2) by

$$g(q_x) = f(x) \qquad (3.3)$$

and $$g(\mu_x) = f(x) \qquad (3.4)$$

which indicates that the likelihood function depends in some smooth way on x. Note that $f(x)$ could be a parametric model, as before, so that this framework includes generalized linear models. However, $f(x)$ is, in general, an unspecified smooth function. Note also that this is the simplest form of model: it is possible to introduce other factors into the model, such as duration. In this case, the function is bivariate (see for example Currie and Waters, 1991 and Renshaw and

Haberman, 1996). The aim of the graduation is to obtain an estimate of *f(x)*, $\hat{f}(x)$. There is a wide range of methods for obtaining $\hat{f}(x)$, known in general as smoothers. We describe below the smoothers that are already used in graduation theory, together with some other useful smoothers which have not yet been applied in graduation. Note that the non-parametric smoothers in the actuarial literature are usually applied to the crude rates, using a normal approximation to the exact likelihood. This is unnecessary, and is more complicated (since the variance of the crude rates must be estimated).

A full description of generalised additive models and smoothers, together with details of the estimation algorithms (which is again an iterative scoring algorithm), can be found in Hastie and Tibshirani (1990). However, these models can again be easily fitted within a statistical package, S-Plus. S-Plus can also estimate generalized linear models, and hence it is possible to perform both parametric and non-parametric graduation within a single package. This gives an extremely useful unified framework for graduation. All graduations (within the scope of the models given by equations (3.1) to (3.4)) can be performed within the same package and their results compared.

The remaining subsections of this section of the paper will describe some of the useful forms for the non-parametric smoothing. Thus, the method of obtaining $\hat{f}(x)$ will be discussed in more detail. In order to describe the smoothers as straightforwardly and clearly as possible, we will consider first the simple regression model. Suppose we have a set of observations of a response variable *Y* and a single predictor *X*, and we wish to obtain an estimate of

$$E(Y|x) = f(x). \tag{3.5}$$

The aim is to obtain an estimate $\hat{f}(x)$ of the arbitrary unspecified smooth function *f(x)*. In relation to the models described in section 2 for graduation, this model is simpler because we will assume that the data are homoscedastic, and use an identity link function. The smoothing procedures are described below for the simple model, and have then to be inserted into the numerical optimisation algorithms for the generalised linear models described in section 2.

3.2 Bin Smoothers

This is a very simple method for smoothing, and is unlikely to be useful in graduation. However, it is described here as an introduction to smoothers. The estimate $\hat{f}(x)$ is obtained by finding the average value of *Y* in each of a number of disjoint and exhaustive regions for the predictor variable. *i.e.* choose a number of cutpoints $c_0 < c_1 < \ldots < c_K$ with $c_0 = -\infty$ and $c_K = \infty$. Then calculate the average of Y in each region. Figure 3.1 shows the results of applying this procedure to the data from table 17.4 of Forfar *et al* (1988). The regions have been chosen to contain 8 points each: (up to 25), (26-33), (34-41), etc. The graduation is affected by the widths of the regions and the starting points. As can be seen from Figure 3.1, the resulting estimate is not very smooth. The problem is caused by the sudden changes from one region to the next. The more successful smoothers described below do not change abruptly like this, but this smoother illustrates an obvious difference from parametric graduation in that the estimate in each region uses only the data in that region and not the complete data set.

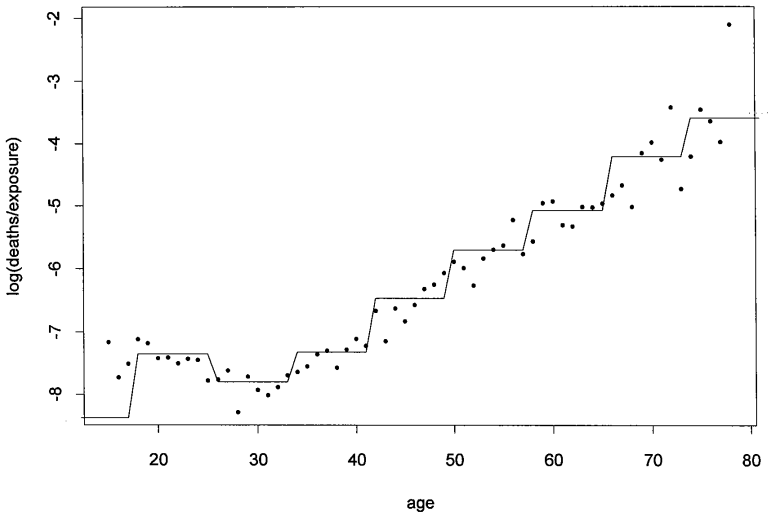

Figure 3.1 Bin Smoother.

We could make this more satisfactory by increasing the number of regions. This is possible for all smoothers: there is always a way of changing the smoother to alter the smoothness and goodness-of-fit of the graduation. The resulting graduation would be a better fit, but would be less smooth than the graduation in Figure 3.1. This illustrates the fundamental trade-off between smoothness and goodness-of-fit, or between variance and bias.

No matter what regions we choose, this method will not give a satisfactory graduation, and we need to consider more sophisticated smoothers.

3.3 Whittaker Graduation and Cubic Smoothing Splines

Whittaker graduation specifies the trade-off between smoothness and goodness-of-fit explicitly by specifying the function below to be minimised:

$$\sum\left(y - f(x)\right)^2 + \lambda \sum\left(\nabla^2 f(x)\right)^2 \qquad (3.6)$$

where $\nabla^2 f(x)$ indicates second differences in the estimates.

i.e. $\nabla^2 f(x) = f(x) - 2f(x-1) + f(x-2)$.

Note that we have assumed here that the values of x are equally spaced (as in graduation), although this is not necessary. It can be shown (Reinsch 1967) that the estimate $\hat{f}(x)$ is a natural cubic spline with knots at each value of x (also known as a smoothing spline). The exact results

of the usual Whittaker graduation as described in the actuarial literature can be obtained by using a normal approximation to the crude rates (and estimating the variances). However, it is argued by Verrall(1993) that this is inconvenient and unnecessary and that it is better to use the appropriate sampling distribution for A_x. This can be done by applying a smoothing spline within the framework of generalised additive models. Figure 3.3 shows a typical such graduation.

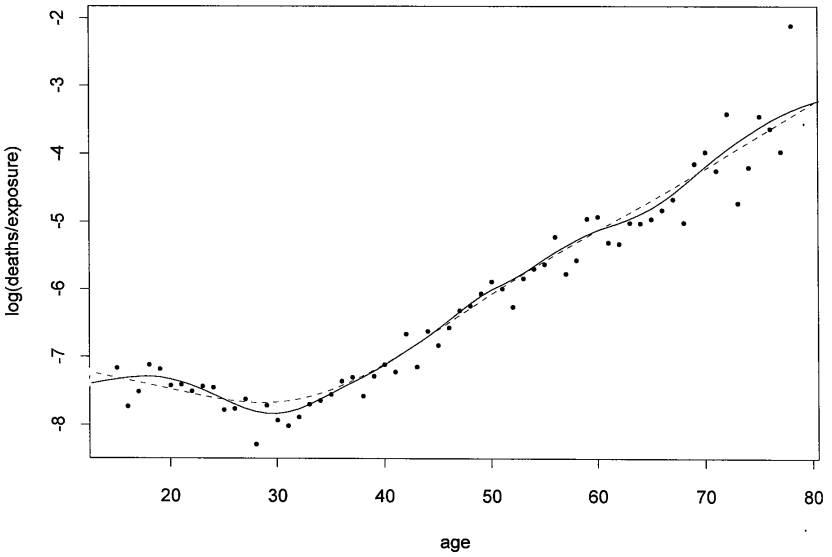

Figure 3.3 Whittaker graduation with 4 degrees of freedom (---), and 10 degrees of freedom (—).

Figure 3.3 shows the result of applying cubic smoothing spline smoothers to the data from table 17.4 of Forfar *et al* (1988). It can be seen that the greater the model degrees of freedom, the less smooth is the graduation.

Note that we are able to use the exact distribution for A_x , and we replace the linear predictor η_x in a GLM by a cubic smoothing spline smoother. In order to reproduce exactly the results of the usual Whittaker method, it is necessary to model the crude rates ($\overset{o}{q}_x$ or $\overset{o}{\mu}_x$) using a normal approximation. For example, for graduating q_x, the model would be

$$\overset{o}{q}_x \sim approx.\ N\left(q_x, \frac{q_x(1-q_x)}{R_x}\right).$$

Thus, $\overset{o}{q}_x$ would be defined as the response variable and weights would have to be defined (and estimated) to account for the variance of $\overset{o}{q}_x$. A cubic smoothing spline used to smooth the crude weights would then reproduce exactly the results from a Whittaker graduation and would minimise $\sum\left(\overset{o}{q}_x - q_x\right)^2 + \lambda\sum\left(\nabla^2 q_x\right)^2$. However, it is inconvenient and unnecessary to resort to a normal approximation. It is better to use the exact distribution of A_x and apply the cubic smoothing spline within a generalised additive model.
The parameter λ governs the trade-off between goodness-of-fit and smoothness (bias and variance), and can be chosen subjectively. Alternatively, an automatic method, such as cross-validation, may be used.

3.4 Moving-Weighted Averages and Kernel Graduation.

Moving-Weighted-Averages (MWA) have a long history in graduation, one of the most famous examples being Spencer's 21 point formula (Spencer, 1904):

$$\hat{q}_x = \frac{1}{350}\Big[60\overset{o}{q}_x + 57\left(\overset{o}{q}_{x+1} + \overset{o}{q}_{x-1}\right) + 47\left(\overset{o}{q}_{x+2} + \overset{o}{q}_{x-2}\right) + 33\left(\overset{o}{q}_{x+3} + \overset{o}{q}_{x-3}\right) + 18\left(\overset{o}{q}_{x+4} + \overset{o}{q}_{x-4}\right) + 6\left(\overset{o}{q}_{x+5} + \overset{o}{q}_{x-5}\right)$$

$$-2\left(\overset{o}{q}_{x+6} + \overset{o}{q}_{x-6}\right) - 5\left(\overset{o}{q}_{x+7} + \overset{o}{q}_{x-7}\right) - 5\left(\overset{o}{q}_{x+8} + \overset{o}{q}_{x-8}\right) - 3\left(\overset{o}{q}_{x+9} + \overset{o}{q}_{x-9}\right) - \left(\overset{o}{q}_{x+10} + \overset{o}{q}_{x-10}\right)\Big]$$

For regression problems kernel smoothers were first suggested by Watson (1964) and Nadaraya (1964). In the actuarial literature, kernel smoothing methods were first used by Copas and Haberman (1983) and Ramlau-Hansen (1983). The connection between kernel smoothing and MWA graduation was expounded in detail by Gavin *et al* (1993), and further aspects and extensions were explored in Gavin *et al* (1994, 1995). These last three papers, which contain some illustrations of the methods, are recommended as an introduction to the subject. More general texts which give introductions to kernel smoothing in a statistical setting are Silverman (1986), Scott (1992) and Wand and Jones (1995).

Consider a set of crude mortality rates, $\overset{o}{q}_i$ for each age x_i ($i = 1, \ldots, n$). The Nadaraya-Watson (N-W) kernel estimate is defined as

$$\hat{q}_x = \frac{\sum_{i=1}^{n} \overset{o}{q}_i\, K_b(x - x_i)}{\sum_{i=1}^{n} K_b(x - x_i)} \qquad (3.7)$$

where $K_b(x) = b^{-1}K(x/b)$, and K is a kernel function. A kernel function must satisfy $\int_{-\infty}^{\infty} K(x)dx = 1$. Note that any probability density function may serve as a kernel function, and a commonly used kernel function is the standard normal density function. However, there is no requirement that the kernel function should be non-negative, and in fact the optimal smoothing kernels, such as the Epanechnikov kernel (Epanechnikov, 1969) can take negative values (see Gavin *et al,* 1993).

By rewriting (3.7) as

$$\hat{q}_x = \sum_{i=1}^{n} S_{xi} \overset{o}{q}_i \qquad (3.8)$$

where $S_{xi} = \dfrac{K_b(x - x_i)}{\sum_{i=1}^{n} K_b(x - x_i)}$,

it can be seen that kernel graduation is very similar to MWA graduation. This is explained in greater detail in Gavin *et al* (1993), where the relationship between Spencer's formula and optimal smoothing kernels is also explained.

The parameter *b,* which is known as the bandwidth, governs the amount of smoothing in the graduation process, in much the same way as λ in Whittaker graduation (see section 3.3). It is well-known from the statistical literature (see Gavin *et al*, 1994, for a discussion of this) that the choice of bandwidth is of much greater significance than the choice of the kernel function. This is somewhat at variance with the actuarial literature, in which much attention has been paid to the choice of kernel function (or equivalently to the form of the weights in MWA graduation). The bandwidth may be chosen subjectively or by a data-driven method such as cross-validation (Stone, 1974). The cross-validation bandwidth minimises

$$\frac{1}{n}\sum_{i=1}^{n}\left(\overset{o}{q}_i - (\hat{q}_i)^{(-i)}\right)^2$$

where $(\hat{q}_i)^{(-i)}$ is the estimated value at x_i calculated without using the data at that age.

There is evidence from the statistical literature that the cross-validation bandwidth may under-smooth the data (see for example Scott, 1992), and this will be of even greater significance in graduation since actuaries generally require a higher degree of smoothness. Scott (1992) has a good discussion of the methods of selecting a value for *b*. For a description of cross-validation in the context of graduation, see Gavin *et al* (1994).

In the examples in this paper, the standard normal density will be used as the kernel function.

Copas and Haberman (1983) suggested an alternative form of kernel graduation:

$$\hat{q}_x^{CH} = \frac{\sum_{i=1}^{n} A_i K_b(x - x_i)}{\sum_{i=1}^{n} R_i K_b(x - x_i)} \qquad (3.9)$$

This can be regarded as the estimator which maximises the local binomial log likelihood using a kernel-weighted likelihood approach (see Staniswalis, 1989 and Copas, 1983). Gavin *et al* (1994) section 2.3 compare the two approaches, (3.7) and (3.9) and give reasons why (3.7) is likely to have a superior performance. In many cases, a more satisfactory graduation can be obtained by first transforming the crude rates using, for example, the logit, Weibull, Gompertz or other transformation. The aim is to find transformed crude rates which broadly follow a straight line, so that the bias inherent in any graduation is reduced over much of the range.
Thus, the usual form of kernel (or MWA) graduation is

$$\hat{q}_x^{t} = \frac{\sum_{i=1}^{n} q_x^{o\,t} K_b(x - x_i)}{\sum_{i=1}^{n} K_b(x - x_i)} \qquad (3.10)$$

where $K(x)$ is the standard normal density function, and t denotes a transformation.

At the boundaries of the graduation (low or high ages), the kernel function increasingly overlaps the end of the table, which results in increased bias. This problem has been studied in detail in the actuarial literature for MWA graduation (see Greville 1981a,b,c and Hoem and Linneman 1988). Gavin *et al* (1995) show how a boundary-correcting kernel suggested by Jones (1993) can be used to overcome problems in the tails of the graduation.
Gavin *et al* (1995) also studied the use of adaptive kernel estimates. The idea underlying these is to allow the bandwidth (and hence the amount of smoothing) to change over the range of the graduation to reflect the exposure. Thus, where the amount of exposure is large, the crude mortality rate will be more reliable, and a lower value for b is used. This results in an estimate which more closely reflects the crude rate. When the exposure is small, b will be larger which results in a greater degree of smoothing being applied. The results from both the above refinements are quite promising and are reported in detail in Gavin *et al* (1995).
Local polynomial modelling (see Fan and Gijbels, 1996) overcomes the boundary problems of graduation in an extremely elegant way. An example of this type of graduation is given in the following section.

3.5 Local Polynomial Graduation.

There are a number of other non-parametric smoothers available which are potentially very useful for graduation and which can be applied within the general framework set out above. Of these, a method which fits low-order polynomials locally appears to be the most promising. The "locally-weighted regression smoother" (*loess*) (Cleveland, 1979), is described below. Since this method fits (using a local smoothing method) a straight line, or a quadratic instead of a constant,

it has the advantage of automatically adjusting for bias at the boundaries. This feature of local polynomial smoothers is discussed in detail in Fan and Gijbels (1996).
We define loess by considering the case of a single predictor, X, for a normally distributed response variable Y (as in simple linear regression). It can be used in the context of generalized additive models as outlined in section 3.1. Consider a sample of size m:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}.$$

The fitted values at some point x_0, $\hat{f}(x_0)$, is an estimate of the underlying smooth trend in Y and for loess, $\hat{f}(x_0)$ is obtained as follows.

Define $N(x_0)$ to be the set of k nearest neighbours of x_0. k may also be defined by deciding the percentage of the data points to include in $N(x_0)$: this is called the *span*.

Calculate $\Delta(x_0) = \max_{x_i \in N(x_0)} |x_0 - x_i|$.

Calculate weights, w_i, for each point in $N(x_0)$, where

$$w_i = T\left(\frac{|x_0 - x_i|}{\Delta(x_0)}\right)$$

and T is the tri-cube weight function:

$$T(u) = \begin{cases} (1-u^3)^3 & \text{for } 0 \le u \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Regress Y on X in $N(x_0)$ using the weights $\{w_1, w_2, \ldots, w_m\}$. This regression can be linear, quadratic, cubic, etc. The fitted value at x_0 is $\hat{f}(x_0)$.

The default setting in S-PLUS (see the example in section 4) is a local linear regression, which may not be satisfactory in the context of graduation. It can be seen that this is similar to a standard regression, except that the contribution of the points to the fit at x_0 depend on their distance from x_0. In practice, it may be better to use a local quadratic regression, by changing the degree of the curve in the *loess* command.
This method of graduation has some similarities with the dynamic generalized linear models suggested by Verrall (1993), and with Whittaker graduation, which can be regarded as a straight line model whose gradient varies smoothly with age (see Verrall, 1994).

4. Implementation in S-PLUS.

The generalized linear models and the generalized additive models can all be implemented in S-PLUS. This is a very flexible environment which combines an object-oriented programming language, a diverse and coherent collection of tools for statistical analysis and extensive graphical facilities. This makes it highly suitable for graduation, and it is possible to design a graduation programme which fits, evaluates and compares all the above methods of graduation. Books which contain detailed descriptions of the S-PLUS system and its use in statistics include Chambers and Hastie (1992) and Venables and Ripley (1994). The purpose of this paper is to emphasise the possibility of introducing a significant amount of coherence to graduation techniques. It also makes it possible to use non-parametric graduation methods in conjunction with parametric methods. For example, there is a developing field of research into the use of non-parametric smoothing methods to assess and improve parametric models. One way to do this is to fit a parametric model and then smooth the residuals using a non-parametric smoother. This was briefly considered in Gavin *et al* (1995) and will be investigated further in section 5.
As an example of how S-PLUS can be used the code is given below for carrying out a parametric straight line regression graduation and a Whittaker graduation (using the exact likelihood rather than the normal approximation usually used in the actuarial literature). It is assumed that q_x is being graduated, using a binomial distribution and a logit link function. The numbers of deaths, the exposure and the ages are assumed to be in the vectors `deaths`, `exposure` and `age` respectively.
We illustrate the results using the data from table 17.4 of Forfar *et al* (1988).

```
crude.rate <- deaths/exposure
binmat <- cbind(deaths, alive = exposure - deaths)
g.glm <- glm(binmat ~ age, family = binomial)
g.whit <- gam(binmat ~ s(age), family = binomial)
win.graph()
par(mfrow = c(1, 2))
plot(age, crude.rate, main = "Straight Line")
lines(age, fitted.values(g.glm))
plot(age, crude.rate, main = "Parametric Graduation")
lines(age, fitted.values(g.whit))
```

The function `glm` fits a generalized linear model, in this case the distribution is specified to be binomial using the `family=` option. In S-PLUS, binomially distributed data is handled by defining a matrix whose first column contains the number of successes (deaths) and whose second column contains the number of failures (alive=exposure-deaths). This is the reason for the `cbind` command which binds together the vectors column-wise. It is straightforward to use other linear predictors. For example, a quadratic polynomial could be fitted (using orthogonal polynomials) by the command
`g.glm <- glm(binmat ~ poly(age,2), family = binomial).`
The results of this fit are assigned to an object, which is here called `g.glm`. This object can be used in a number of ways. The commands above show how a plot of the graduation can be obtained (the result of this is shown below). The command `summary(g.glm)` gives the following information:

```
Call: glm(formula = binmat ~ age, family = binomial)
Deviance Residuals:
       Min         1Q     Median         3Q      Max
 -3.507917 -1.007146 0.02162252 0.7828463 2.334852

Coefficients:
                Value   Std. Error   t value
(Intercept) -8.87182491 0.0596038189 -148.8466
        age  0.08428521 0.0008177627  103.0681

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 9865.668 on 49 degrees of freedom

Residual Deviance: 84.95456 on 48 degrees of freedom

Number of Fisher Scoring Iterations: 3

Correlation of Coefficients:
    (Intercept)
age -0.9966998
```

Some explanation is required for this output. The deviance residuals are defined in section 2.4 of McCullagh and Nelder (1989), and are residuals defined to be suitable for generalized linear models. The lowest, highest, lower and upper quartiles and median of these is given. The parameter estimates are called coefficients in S-PLUS. The deviance (see section 2.3 of McCullagh and Nelder, 1989) measures the goodness-of-fit and can be used for comparing models. Model choice is a balance between the deviance and the degrees of freedom (goodness-of-fit versus smoothness). The routine used to fit the models is called a Fisher scoring iteration. To use Whittaker graduation, all that is necessary is to change from `glm` to `gam` and to use the cubic smoothing spline smoother, `s()`. The command `win.graph()` opens a graphics window, `par(mfrow = c(1, 2))` divides this window into two so that two plots can be shown together, and `plot()` and `lines()` are generic functions which plot the crude rates and draw lines for the fitted rates. The graphical output is shown in Figure 4.1.

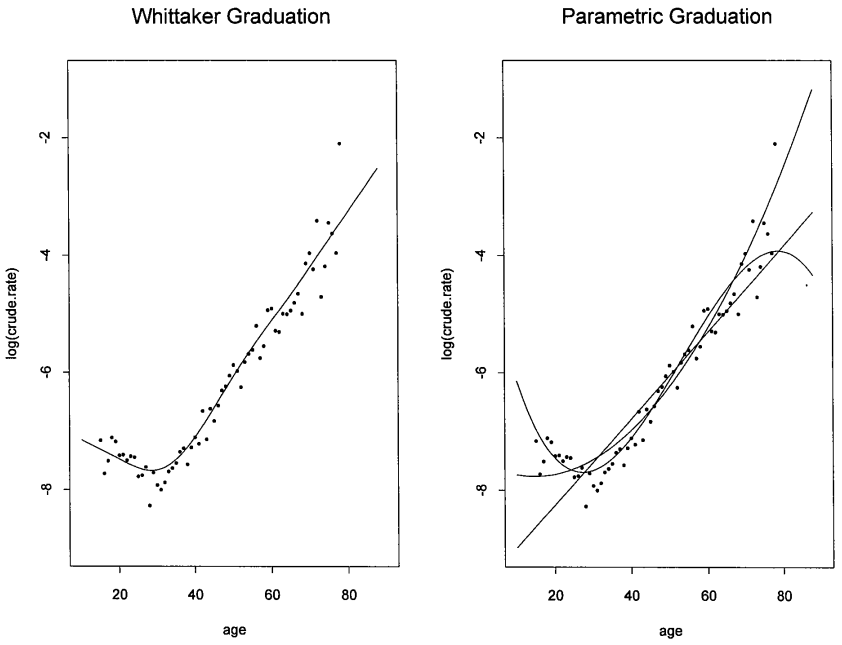

Figure 4.1 Whittaker graduation with 4 degrees of freedom and straight line, quadratic and cubic parametric fits.

To use a Poisson distribution for mortality rates, the following commands are appropriate:

```
g.glm <- glm(deaths ~ age + offset(log(exposure)), family =
poisson)
g.whit <- gam(deaths ~ s(age) + offset(log(exposure)), family =
poisson)
```

The locally-weighted regression smoother (loess) can be applied by replacing `s(age)` by `lo(age)` in the `gam` functions. Thus, a loess smooth applied to the same data as above would be obtained from

```
g.loess <- gam(deaths ~ lo(age) + offset(log(exposure)), family =
poisson)
```

The loess smoother is used in the example below which illustrates non-linear graduation.

To implement a non-linear graduation, the likelihood function itself must be specified explicitly in order that the relevant calculations can be made. Consider first graduating q_x using the LGM*(r,s)* function. The following function calculates the likelihood for this model, where `gm` is the non-linear predictor of the form

$$\sum_{i=0}^{r-1}\alpha_i x^i + \exp\left(\sum_{i=0}^{s-1}\beta_i x^i\right).$$

```
lbin <- function(gm,deaths,exposure)
          exposure*log(1+gm)-deaths*log(gm)
```

When graduating μ_x using the GM*(r,s)* function, the function below is appropriate.

```
lpois <- function(gm,deaths,exposure)
          exposure*gm-deaths*log(gm)
```

As is well known, great care must be taken with non-linear graduation to ensure that the parameter estimates correspond to a global maximum. Starting values must be supplied which avoid values for `gm` which are not feasible (for example `gm` cannot be negative). It is usual to keep all the data for each graduation together in a dataframe which can be specified in the appropriate functions for generalized linear or additive models. The same applies to non-linear graduation, and in this case the parameters themselves must be explicitly specified in the predictor and starting values included in the dataframe. For this purpose, the dataframe with the parameters attached becomes a parameterised dataframe. We consider the graduation of the data from table 17.4 of Forfar *et al* (1988). The complete sequence of functions for this graduation is given below. It is assumed that the data have been imported into S-PLUS as the data frame `fmw` with the names of the variables in the top line.

```
parameters(fmw) <- list(alpha0=0, alpha1=0,beta0=0,beta1=0)
attach(fmw,1)
age1 <- (age-70)/50
alpha0 <- -0.005
alpha1 <- -0.005
beta0 <- 4
beta1 <- -3.2
detach(1,save="fmwnew")
fit.fmw <-ms(~lpois(alpha0+alpha1*age1+exp(beta0+beta1*age1),
deaths, exposure),fmwnew)
```

This assumes that the function `lpois` has already been created. The object `fit.fmw` contains ... and the fitted values may be obtained as follows.

```
p.fmw <- fit.fmw$parameters
fv.fmw <- p.fmw[1]+ p.fmw[2]*age1+ exp(p.fmw[3]+p.fmw[4]*age1)
```

Thus, the graduation may be plotted using

```
plot(age,log(deaths/exposure))
lines(age,log(fv.fmw))
```

We illustrate this and the loess smoother together. We will use the default settings for the smoothing parameter in loess, and first consider the default smoothing function (locally linear smoother). The plot below shows the loess smoother together with the non-linear graduation used by Forfar *et al.*

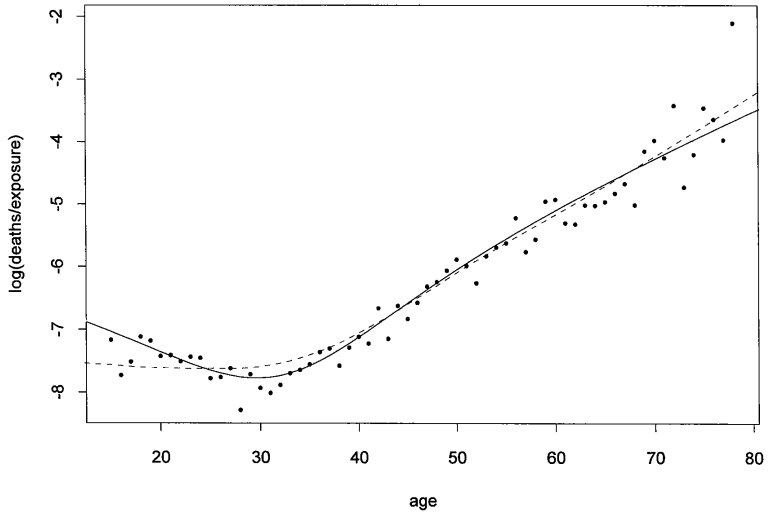

Figure 4.2 Graduation of Forfar *et al* (solid line) and loess graduation (dotted line).

Clearly the loess smoother does not produce a satisfactory graduation since the shape is not captured satisfactorily. It would be possible to experiment with the smoothing parameter. However, the shape of the graduation seems to indicate that a local quadratic smoother would be more appropriate. This can be fitted using the following function.

```
g.loess <- gam(deaths~lo(age,degree=2))
```

The results of this graduation are shown in Figure 4.3.

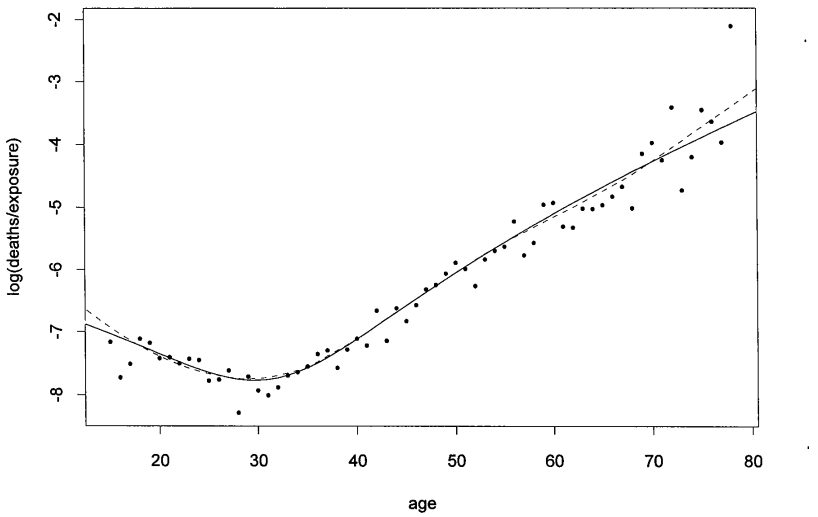

Figure 4.3 Graduation of Forfar *et al* (solid line) and loess graduation (dotted line).

Clearly this is more satisfactory, and it could represent an improvement on the parametric graduation.

5. Practical Implications

This paper has shown how all the standard methods of graduation can be applied within a unified framework. It has also suggested that there are other smoothing procedures which have great potential benefits and represent improvements over existing methodology. However, it has not been the aim to advocate one particular method in preference to any other. The advantage of having parametric and non-parametric graduation techniques all available together is that they can all easily be applied and compared in the process of arriving at a suitable graduation. There are other uses to which the non-parametric smoothing techniques could be put, which are discussed below.

For example, a practitioner may prefer to use a parametric graduation but might find a non-parametric method with its greater inherent flexibility useful for exploring the data to decide the range of parametric models to use. It is possible that a practitioner whose preference was for parametric models might find that a non-parametric graduation was more satisfactory, were it easily available.

Another important advantage of the approach outlined in this paper is in assessing the goodness-of-fit of the graduated values. Naturally, it is possible to use all the standard goodness-of-fit

criteria in S-PLUS (χ^2 tests, runs tests, signs tests, etc.) as suggested by the standard actuarial literature. It is also straightforward to examine residual plots, using the advanced graphical facilities available in S-PLUS. But a further useful procedure which was first suggested by Gavin *et al* (1995) is to apply a non-parametric smoother to the residuals as an aid to the identification of departures from the fitted model.

Finally, it is possible to use parametric and non-parametric graduations together. A simple parametric model might provide fairly reasonable overall graduation, but one which is not completely satisfactory throughout the whole range of ages being graduated. Gavin *et al* (1995) suggested applying a non-parametric graduation to the residuals (suitably defined) from the parametric model and adjusting the parametric fit accordingly.

These new techniques for using non-parametric graduation in conjunction with parametric graduation are under investigation by the author and will be reported more fully in subsequent papers.

References

Chambers, J.M. and Hastie, T.J. (1992) *Statistical Models in S.* Chapman and Hall, London.

Cleveland, W.S. (1979) *Robust Locally-Weighted Regression and Smoothing Scatterplots.* J.Amer. Statist. Ass., **74**, 829-836.

Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression.* Chapman and Hall, London.

Copas, J.B. (1983) *Plotting p against x.* JRSS, Series C, **32**, 25-31.

Copas, J.B. and Haberman, S. (1983) *Non-parametric graduation using kernel methods.* JIA, **110**, 135-56.

Cox, D.R. (1972) *Regression models and life-tables (with discussion).* JRSS, Series B, **34**, 187-220.

Currie, I.D. and Waters, H.R. (1991) *On Modelling Select Mortality.* JIA, **118**, 453-481.

Epanechnikov, V.A. (1969) *Non-parametric Estimation of a Multivariate Density.* Theory of Prob. and its Applications, **14**, 153-158.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Forfar, D.O., McCutcheon, J.J. and Wilkie, A.D. (1988) *On graduation by mathematical formula.* JIA, **115**, 1-149.

Francis, B., Green, M. and Payne, C. (1993) *The Glim System*, Release 4 Manual, Clarendon Press, Oxford.

Gavin, J.B., Haberman, S. and Verrall, R.J. (1993) *Moving Weighted Average Graduation using Kernel Graduation.* Insurance: Mathematics and Economics, **12**, 113-126

Gavin, J.B., Haberman, S. and Verrall, R.J. (1994) *On the Choice of Bandwidth for Kernel Graduation.* JIA, **121**, 119-134

Gavin, J.B., Haberman, S. and Verrall, R.J. (1995) *Variable Kernel Graduation with a Boundary Correction.* Transactions of Society of Actuaries, **XLVII**.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Greville, T.N.E. (1981a) *Moving-Weighted-Average Smoothing Extended to the Extremities of the Data. I Theory.* Scand.Act.J., 38-55.

Greville, T.N.E. (1981a) *Moving-Weighted-Average Smoothing Extended to the Extremities of the Data. II Methods*. Scand.Act.J., 65-81.

Greville, T.N.E. (1981a) *Moving-Weighted-Average Smoothing Extended to the Extremities of the Data. III Stability and Optimal Properties*. J. of Approximation Theory, 43-58.

Haberman, S. and Renshaw, A.E. (1996) *Generalized Linear Models and Actuarial Science*. The Statistician, **45**, 1-31

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman and Hall, London

Hastie, T. and Tibshirani, R. (1993) *Varying Coefficient Models (with discussion)*. JRSS, Series B, **55**, 757-796.

Hoem, J.M. and Linnemann, P. (1988) *The Tails in Moving Average Graduation*. Scand. Act. J., 193-229.

Jones, M.C. (1993) *Simple Boundary Correction for Kernel Density Estimation*. Statistics and Computing, 135-146.

London, D. (1981) *Graduation - the revision of estimates*. ACTEX publications, Winsted and Abington, CT.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd Edition, Chapman and Hall, London.

Nadaraya, E.A. (1964) *On estimating regression*. Theor. Prob. Appl., **9**, 141-142.

Nelder, J.A. and Wedderburn, R.W.M. (1972) *Generalized Linear Models*. JRSS, Series A, **135**, 370-384.

Ramlau-Hansen, H. (1983) *The choice of kernel function in the graduation of counting process intensities*. Scand. Act. J., 165-182.

Reinsch, C. (1967) *Smoothing by Spline Functions*. Numer.Math., **10**, 177-183.

Renshaw, A.E. (1991) *Actuarial Graduation Practice and Generalized Linear and Non-linear Models*. J.I.A., **118**, 295-312.

Renshaw, A.E. (1992) *Joint Modelling for Actuarial Graduation and Duplicate Policies*. J.I.A., **119**, 69-85.

Renshaw, A.E. and Haberman, S. (1995) *On the graduations associated with a multiple state model for permanent health insurance*. Insurance: Mathematics and Economics, **17**, 1-17.

Renshaw, A.E. and Haberman, S. (1996) *Dual Modelling and Select Mortality*. Insurance: Mathematics and Economics, to appear.

Renshaw, A.E., Haberman, S. and Hatzopoulos, P.(1996) *The modelling of recent mortality trends in United Kingdom male assured lives*. British Actuarial Journal, **2**, 449-477.

Renshaw, A.E. and Hatzopoulos, P.(1996). *On the Graduation of Amounts*. British Actuarial Journal, **2**, 185-205.

Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice and Visualisation*. John Wiley and Sons.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Spencer, J. (1904) *On the graduation of the rates of sickness and mortality presented by the experience of the Manchester Unity of Oddfellows during the period 1893-1897*. JIA, **38**, 334-43.

Staniswalis, J.G. (1989) *The kernel estimate of a regression function in likelihood-based models*. J.Amer.Statist.Ass., **84**, 276-283.

Stone, M. (1974) *Cross-validatory choice and the assessment of statistical predictions (with discussion)*. JRSS, Series B, **36**, 111-147.

Venables, W.N. and Ripley, B.D. (1994) *Modern Applied Statistics with S-PLUS*. Springer-Verlag.

Verrall, R.J. (1993) *Graduation by Dynamic Regression Methods*. JIA, **120**, 153-170.

Verrall, R.J. (1994) *Whittaker graduation and dynamic generalized linear models*. Insurance: Mathematics and Economics, **13**, 7-14.

Verrall, R.J. (1995) *Whittaker graduation and Parametric State Space Models*. Actuarial Report Number 79, Department of Actuarial Science and Statistics, City University.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. Chapman and Hall, London.

Watson, G.S. (1964) *Smooth regression analysis*. Sankhya, A(**26**), 359-372.

Whittaker, E.T. (1923) *On a new method of graduation*. Proc. Edinburgh Math. Soc., **41**, 63-75.